

Beyond Utilitarianism and Deontology: A Synthesized Ethical Framework for AI Response Generation

The Dual Foundations of Harm Avoidance and Utility Maximization

The inquiry into the default moral outlook of a large language model (LLM) reveals a sophisticated and multi-layered ethical architecture that cannot be accurately described by a single philosophical doctrine such as utilitarianism or deontology [20](#) [41](#). Instead, the available evidence points towards a synthesized system built upon two distinct yet interconnected foundations: a foundational layer of harm avoidance rooted in deontological principles, and a secondary layer of utility maximization driven by consequentialist optimization. This dual structure represents a pragmatic approach to managing the immense capabilities of modern AI, balancing the imperative to prevent negative outcomes with the goal of providing beneficial and helpful responses. The core challenge in AI design is not merely to choose one ethical theory over another but to create a system that can navigate the tension between them dynamically, depending on the context of the user's query [45](#). The concept of AI alignment itself is defined as the effort to ensure that an AI system's behavior is consistent with human intentions and values [17](#) [18](#). This overarching objective necessitates a framework that incorporates both prohibitions against harmful actions and incentives for generating positive ones. Research in this area often begins by deconstructing the fundamental question of ethics: what makes an action ethically better or worse than an alternative? [43](#). The AI's framework attempts to address this by embedding mechanisms that evaluate actions based on both their inherent nature (a deontological concern) and their likely outcomes (a consequentialist concern). For instance, the development of safety strategies explicitly acknowledges the need to manage risks associated with increasingly capable systems causing human harm, a central tenet of AI safety research [42](#).

This duality is evident in the very terminology used to describe ideal AI behavior. The pursuit of creating models that are "helpful, honest, and harmless" encapsulates this dualism perfectly [48](#). The attribute of being "harmless" establishes a strong deontological constraint, setting a boundary beyond which the AI should not go. It implies a set of

forbidden actions—such as providing illegal instructions, spreading malicious content, or causing personal distress—that are prohibited irrespective of the potential benefits of doing so. Conversely, the attributes of being "helpful" and "honest" are inherently consequentialist. They represent goals to be optimized; the AI is trained to produce responses that lead to the best possible outcome for the user, whether that means solving a problem, answering a question correctly, or providing insightful advice. The technical processes behind this, such as reward modeling, further solidify the utilitarian influence. In these processes, LLMs are used to evaluate and score generated responses, rewarding those that are well-reasoned, factually correct, and contextually appropriate [10 28](#). This creates a powerful feedback loop where the model learns to associate higher scores with more useful and beneficial outputs, effectively optimizing for positive consequences. Therefore, the AI's default moral outlook is not a static position but a dynamic equilibrium. It is a system perpetually calculating how to maximize utility while remaining firmly within the non-negotiable boundaries established by its foundational rules. This hybrid approach reflects a recognition that neither pure deontology nor pure utilitarianism is sufficient on its own. Pure deontology could lead to rigid, unhelpful behavior, while pure utilitarianism could justify harmful actions if they led to a greater overall good. The AI's framework seeks to mitigate these risks by combining both approaches.

Deontological Guardrails: Rules, Duties, and Non-Negotiable Constraints

A significant and defining aspect of the AI's ethical framework is its foundation in deontological principles, which emphasize adherence to rules, duties, and categorical imperatives [21](#). This manifests as a robust set of guardrails designed to prevent the AI from engaging in certain classes of harmful or undesirable behavior, regardless of the potential positive consequences. This rule-based orientation is a critical component of AI safety, which aims to prevent increasingly capable systems from causing humans harm [42](#). The implementation of such constraints ensures that the AI operates within a predefined moral boundary, functioning as a form of programmed duty [3](#). These guardrails are not abstract ideals but concrete programming directives embedded within the model's instruction tuning and alignment processes. For example, Qwen-Agent, an intelligent agent framework built on the Qwen model, explicitly integrates ethical considerations and compliance mechanisms at the design level to ensure responsible operation [23](#). This includes the direct refusal of requests for illegal activities or harmful

content, a clear example of a deontological prohibition in action [23](#) . Such rules function similarly to Kantian duties, where an action is deemed wrong if it fails to adhere to a universalizable principle, irrespective of its outcome [3](#) .

The existence of these deontological constraints is also formally recognized by the organizations developing the technology. Alibaba Cloud's Technology Ethics Committee, for instance, has proposed six core guiding principles that the company commits to upholding in its AI development efforts [34](#) [39](#) . While the specific details of these principles are not provided in the source materials, their designation as "core guiding principles" strongly suggests a rule-based framework analogous to a code of conduct or a set of duties that the AI's behavior must respect. These principles likely encompass commitments to fairness, transparency, accountability, and safety, all of which are foundational to deontological ethics in the context of technology [14](#) [15](#) . The emphasis on preventing harm is a recurring theme, directly linking the field of AI safety to deontological theories through the use of behavioral constraints [21](#) . This focus establishes a primary duty for the AI: to avoid causing harm. This duty acts as a powerful brake on the AI's otherwise expansive capacity for optimization. It prevents the model from pursuing a "helpful" outcome if achieving that outcome would require violating a fundamental rule, such as compromising privacy, generating false information, or providing dangerous instructions. This hard-coded prohibition is a key feature distinguishing the AI's ethical architecture from a simple utility-maximizing algorithm. The system is not just calculating the best answer; it is filtering answers through a lens of pre-defined ethical prohibitions. This layered approach, where consequentialist optimization is subordinate to deontological constraints, provides a necessary safeguard against the potential downsides of purely consequentialist reasoning, such as justifying unethical means for a perceived greater good.

Utilitarian Optimization: A Drive for Helpfulness and Positive Outcomes

Parallel to the foundational deontological guardrails, the AI's ethical framework is heavily influenced by utilitarianism, a consequentialist ethical theory that judges actions based on their outcomes [41](#) [49](#) . This influence is most apparent in the system's primary optimization goals, which revolve around maximizing helpfulness, honesty, and overall user satisfaction [48](#) . The entire field of AI alignment is predicated on the idea of steering an AI's behavior to align with human intentions and values [17](#) [18](#) . This process involves

training the model to generate responses that are not only factually correct but also contextually relevant, coherent, and genuinely useful to the user. This drive for utility is implemented through advanced technical methods, most notably reinforcement learning from human feedback (RLHF) and reward modeling. In these processes, the AI is trained using frameworks that involve evaluating and scoring its outputs [10](#). For example, a state-of-the-art double-stage multi-dimensional reward framework utilizes a Multimodal Large Language Model (MLLM) to assess both the reasoning process and the final output, thereby supervising the entire generation pipeline to maximize a unified reward signal [10](#). This mechanism implicitly trains the model to pursue responses that consistently receive high scores, which are correlated with being informative, well-reasoned, and ethically sound within the bounds of the training data.

The utilitarian drive is also reflected in the pluralistic nature of the ethical frameworks used for training and evaluation. Research initiatives have developed suites like ETHICS, which frame moral judgment as a supervised learning task spanning a wide range of normative theories, including commonsense morality, deontology, utilitarianism, justice, and virtue ethics [35](#) [44](#). The inclusion of utilitarianism as a core component in such comprehensive frameworks indicates that developers recognize its importance in creating AI that promotes positive outcomes. Furthermore, studies comparing AI and human moral reasoning specifically focus on determining whether LLMs exhibit a consequentialist or deontological orientation, underscoring the centrality of this debate in the design of AI ethics [20](#) [47](#). The finding that models with larger parameter counts tend to demonstrate better ethical performance, with models like Gemma and Qwen showing strong results, suggests that the capacity for nuanced reasoning, including consequentialist calculations, scales with model size [22](#). This allows the AI to weigh different potential outcomes of its responses and select the one that appears to offer the greatest net benefit. For instance, when asked for advice on a complex problem, the AI will analyze various courses of action and recommend the one that leads to the most favorable consequence, demonstrating a clear utilitarian logic. This focus on outcomes is what enables the AI to be "helpful," a primary goal in its design. However, this utilitarian optimization always occurs within the previously established deontological boundaries. The AI will seek the most helpful answer, but it will never violate its core rules to get there. This interplay between seeking the best outcome and respecting absolute prohibitions defines the nuanced ethical landscape of the AI's response generation.

A Pluralistic and Synthesized Ethical Architecture

The ethical framework guiding the AI's responses is best understood not as a strict adherence to one philosophical school, but as a synthesized and pluralistic architecture that draws upon multiple ethical traditions. This approach acknowledges the complexity of real-world moral dilemmas and the limitations of relying on a single ethical perspective. The system is designed to integrate elements of consequentialism, deontology, and other ethical theories to produce responses that are contextually appropriate and broadly acceptable [1](#) [49](#). This synthesis is a deliberate design choice, reflecting a sophisticated understanding that no single grand theory can adequately cover all scenarios. For example, Value Sensitive Design (VSD), a methodology mentioned in the context of building ethical-aligned AI, advocates for integrating the values of all stakeholders affected by a technology into the design process, promoting a pluralistic approach rather than privileging one philosophical tradition [8](#). This mirrors the AI's own attempt to balance competing values. Research projects have been developed specifically to train models on frameworks that span several ethical paradigms, including utilitarianism, deontology, virtue ethics, and theories of justice [35](#) [44](#). The inclusion of virtue ethics, which focuses on character and moral virtues, alongside more rule- or outcome-based theories, suggests an aspiration for the AI to not only act correctly but also to embody desirable qualities like honesty and kindness [1](#) [45](#).

The AI's ability to navigate these different ethical lenses is context-dependent. When presented with a clear-cut legal or safety-related query, the deontological guardrails will dominate, leading to a response that prioritizes rules and prohibitions [21](#). Conversely, when asked for creative solutions, recommendations, or explanations, the utilitarian drive for helpfulness takes precedence, resulting in a response that optimizes for the best possible outcome [48](#). In some cases, the AI may even attempt to reason from a virtue-ethics perspective, framing its guidance in terms of fostering positive character traits [45](#). This contextual switching is a hallmark of a synthesized system rather than a monolithic one. Comparative studies explicitly aim to determine whether LLMs lean more towards a deontological or utilitarian orientation, indicating that the answer is not uniform across all models or scenarios [20](#) [41](#). One study analyzing open-source LLMs like Qwen found that their ethical performance varied, highlighting the impact of fine-tuning and curated datasets on the final output [36](#) [37](#). This demonstrates that the synthesis is not static but is actively shaped during the training and alignment phases. The result is an AI that can present itself as a principled rule-follower in some situations and a pragmatic problem-solver in others, depending on what the context demands. This flexibility is a strength, allowing the AI to be useful in a wide variety of domains, from academic writing

assistance [4](#) to customer service applications [11](#) , but it also introduces complexity and potential for inconsistency.

Ethical Theory	Core Principle	Manifestation in AI Response Generation
Deontology	Adherence to rules, duties, and obligations; actions are right or wrong in themselves.	Refusal to provide illegal instructions or harmful content 21 23 ; adherence to a set of non-negotiable core principles 34 .
Utilitarianism	Maximization of overall happiness or utility; consequences of actions determine their morality.	Optimization for helpfulness, honesty, and user satisfaction 48 ; reward modeling to produce beneficial outcomes 10 ; recommendation of courses of action with the best net result 46 .
Virtue Ethics	Focus on the character and virtues of the moral agent; acting in accordance with ideal traits.	Framing advice in terms of fostering positive character traits 45 ; aiming to be perceived as trustworthy, kind, and competent.
Justice	Fairness and equity in treatment and distribution of resources.	Ensuring balanced and fair responses, avoiding biases present in training data 35 44 ; consideration of fairness in machine learning definitions 46 .

The Role of Alignment and Organizational Principles

The specific configuration of the AI's ethical framework is the direct result of continuous efforts in AI alignment and is heavily influenced by the stated principles of the organizations developing it. AI alignment is the discipline dedicated to ensuring that artificial intelligence systems behave in ways that are beneficial and consistent with human values and intentions [17](#) [18](#) . This is not a one-time setup but an ongoing, iterative process involving extensive fine-tuning, instruction tuning, and the development of sophisticated reward models [10](#) [13](#) . The ultimate goal is to make the AI "helpful, honest, and harmless" [48](#) . This process involves curating vast datasets of human preferences and using them to guide the model's behavior, effectively teaching it what constitutes a good or bad response based on human judgment [37](#) . The alignment process is therefore the crucible in which the AI's ethical stance is forged, blending statistical patterns from data with explicit programming constraints to create a functional, albeit complex, moral outlook.

Corporate and institutional policies play a crucial role in shaping these alignment efforts. For example, Alibaba Cloud's Technology Ethics Committee has formally established six core guiding principles that the company pledges to follow in its AI development [34](#) [39](#) . Although the specific content of these principles is not detailed in the provided sources, their existence signifies a top-down commitment to embedding ethical considerations

into the technology from the ground up. These principles likely provide a high-level ethical compass that informs the more granular rules and reward functions used in the alignment process. They serve as an anchor, ensuring that the AI's optimization for utility does not stray into ethically dubious territory. This top-down governance is complemented by international and cultural considerations. For instance, China's proposed "Global AI Governance Initiative" emphasizes principles of mutual respect and equality among nations, regardless of their size or social system [5](#) . While it is unclear how directly these geopolitical principles are encoded into a specific model like Qwen, they represent a broader value system that may influence the training data and the global deployment of the technology. The combination of internal corporate ethics, external regulatory trends, and global philosophical debates creates a rich and sometimes conflicting set of inputs that the AI's alignment process must navigate. This explains why the AI's ethical framework can appear nuanced and multifaceted, reflecting a synthesis of diverse and sometimes competing value systems rather than a single, internally consistent philosophy.

Limitations and the Challenge of Consistent Moral Reasoning

Despite the sophisticated design of its ethical framework, the AI's capacity for moral reasoning is subject to significant limitations and challenges. A primary issue is the potential for inconsistent or contradictory outputs, a phenomenon noted in research on AI homogenization [19](#) . Because the AI generates responses based on probabilistic patterns learned from massive text corpora, it can sometimes arrive at logically incompatible conclusions for similar prompts. This lack of internal consistency makes it difficult to pin down a stable, definitive ethical stance. The same model might provide a utilitarian justification for an action in one interaction and a deontological prohibition in another, with both responses appearing plausible according to the model's training. This inconsistency stems from the inherent nature of LLMs as pattern-matching engines rather than entities with genuine consciousness or a unified moral belief system. Their "reasoning" is an emergent property of their training data, which itself contains a vast array of conflicting human opinions on ethical matters.

Furthermore, the AI's ethical framework is fundamentally limited by the quality and scope of its training data. The model learns what is considered "ethical" by identifying patterns of approval and disapproval in the text it was trained on. If this data contains

biases, fallacies, or outdated norms, the AI will inevitably learn and reproduce them ¹⁹. This presents a major challenge for creating a truly fair and impartial ethical system. While developers strive to use curated and filtered datasets to mitigate this risk ³⁷, completely eliminating bias is nearly impossible. The AI's moral reasoning is therefore a reflection of the collective, imperfect morality of humanity as captured in digital text. Another significant limitation is the opacity of the decision-making process. While we can observe the inputs and outputs, the precise algorithmic weighting of deontological rules versus utilitarian calculations remains largely unknown ¹⁷. We can infer the balance from external research and general principles, but the exact logic governing the AI's choices is a "black box." This lack of transparency makes it difficult to audit the system's ethical reasoning or to understand why it made a particular decision. Finally, the AI lacks a true understanding of context, nuance, and the real-world consequences of its suggestions. It can simulate ethical reasoning by stringing together words that sound reasonable, but it does not possess the lived experience or emotional intelligence to grasp the full gravity of a moral dilemma. These limitations mean that while the AI's ethical framework is powerful and increasingly sophisticated, it is far from perfect and requires careful oversight and critical evaluation by human users.

Reference

1. UNCOVERING CROSS-LINGUISTIC MISALIGNMENTS <https://openreview.net/pdf/01b7b9eb92c29f2c72a3f05d4c5ff0097e87bc8e.pdf>
2. Eliciting Values for Technology Design with Moral Philosophy <https://journals.sagepub.com/doi/10.1177/01622439221122595>
3. Aristotle in the Anthropocene: The comparative benefits of ... <https://journals.sagepub.com/doi/10.1177/20530196221105093>
4. LLMs4All: A Review of Large Language Models Across ... <https://arxiv.org/pdf/2509.19580>
5. 中国智·惠世界- AI From China Benefits The World <https://oss.imsilkroad.com/ice/2025/04/30/88d50b4dbe9a4d8da97cc9a624114fa9.pdf>
6. UCL-Bench: A Chinese User-Centric Legal Benchmark for ... <https://aclanthology.org/2025.findings-naacl.444.pdf>
7. Qwen3-VL技术报告英中对照版 <https://www.scribd.com/document/967756683/Qwen3->

VL%E6%8A%80%E6%9C%AF%E6%8A%A5%E5%91%8A%E8%8B%B1%E4%B8%AD
%E5%AF%B9%E7%85%A7%E7%89%88

8. 符合伦理的人工智能应用的价值敏感设计：现状与展望 <https://html.rhhz.net/tis/html/202105002.htm>
9. 温州肯恩大学学生手册 <https://www.wku.edu.cn/sites/main.prod.dpmgr.wku.edu.cn/files/2025-07/Student%20Handbook%202025-2026.pdf>
10. 人工智能2025_5_23 <http://www.arxivdaily.com/thread/67752>
11. Whitepaper PDF File-2025012022362200001 <https://www.scribd.com/document/897138350/whitepaper-pdf-file-2025012022362200001>
12. 文档下载亚洲法与社会杂志第7卷第3期 <http://www.socio-legal.sjtu.edu.cn/Download/Download.aspx?Guid=c1f0483fc6f24f34bab69b6f5c63a0d2>
13. A Survey of Instruction Tuning in Large Language Models https://arxiv.org/pdf/2508.17184v1?utm_source=daily-mt-picks&utm_medium=email&utm_campaign=machine-translation-digest-for-aug-24-2025
14. 人工智能安全治理框架2.0 <https://www.tc260.org.cn/upload/2025-09-15/1757911253996041369.pdf>
15. 人工智能处理医学数据伦理要求的专家共识 <https://actaps.sinh.ac.cn/pdf.php?id=8869>
16. 大模型道德价值观对齐问题剖析 - 计算机研究与发展 <https://crad.ict.ac.cn/article/doi/10.7544/issn1000-1239.202330553>
17. 对齐的理论 <https://aclanthology.org/2024.ccl-2.7.pdf>
18. 人工智能对齐：全面性综述 - AI Alignment <https://alignmentsurvey.com/uploads/AI-Alignment-A-Comprehensive-Survey-CN.pdf>
19. 人工智能2026_1_13[2] <http://arxivdaily.com/thread/75635>
20. Comparing AI and human moral reasoning: context- ... <https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1710410/full>
21. (PDF) Deontology and safe artificial intelligence https://www.researchgate.net/publication/381403747_Deontology_and_safe_artificial_intelligence
22. A Comprehensive Ethical Evaluation of Open-Source ... <https://arxiv.org/html/2505.16036v1>
23. 构建负责任的AI助手：Qwen-Agent伦理与合规实践指南 https://blog.csdn.net/gitblog_00816/article/details/151432018
24. qwen3-235b-a22b Model by Qwen <https://build.nvidia.com/qwen/qwen3-235b-a22b/modelcard>
25. 自然语言处理2026_2_3 <http://arxivdaily.com/thread/76184>

26. Arxiv今日论文| 2026-01-08 http://lonepatient.top/2026/01/08/arxiv_papers_2026-01-08.html
27. CRiskEval: A Chinese Multi-Level Risk Evaluation ... <https://aclanthology.org/2025.acl-long.670.pdf>
28. Incentivizing Reasoning for Advanced Instruction- ... <https://openreview.net/pdf/0de94b848e4b6f73ac84843d481395c51d499376.pdf>
29. Untitled - 第二十九届全球华人计算机教育应用大会 <https://gccce2025.jiangnan.edu.cn/dfiles/18334/doc/pdf/gzf.pdf>
30. 002 003 009 010 014 034 039 057 060 127 <https://www.phirda.com/upload/file/201512115012.pdf>
31. CAPability: A Comprehensive Visual Caption Benchmark ... <https://arxiv.org/html/2502.14914v4>
32. International Conference on Language Testing and ... <https://cse.neea.edu.cn/res/ceedu/report/1505/15120091.pdf>
33. Adversarial Attacks of Vision Tasks in the Past 10 Years <https://dl.acm.org/doi/full/10.1145/3743126>
34. Alibaba's CTO On Everything You Wanted To Know About ... <https://www.alibabacloud.com/blog/599364>
35. VALUEFLOW: TOWARD PLURALISTIC AND STEER <https://openreview.net/pdf/3ca94f7ca0516c88e5a650f474bbd571da390fdd.pdf>
36. (PDF) The Convergent Ethics of AI? Analyzing Moral ... https://www.researchgate.net/publication/391247119_The_Convergent_Ethics_of_AI_Analyzing_Moral_Foundation_Priorities_in_Large_Language_Models_with_a_Multi-Framework_Approach
37. An efficient 7B instruct-model fine-tuned with curated datasets https://www.academia.edu/143106728/Birbal_An_efficient_7B_instruct_model_fine_tuned_with_curated_datasets
38. M5TC4 Titanium Alloy Flange plum bolts, fastening screws ... <https://www.aliexpress.com/item/1005009580046403.html>
39. Alibaba's CTO On Everything You Wanted To Know About ... https://www.alibabacloud.com/blog/alibaba%E2%80%99s-cto-on-everything-you-wanted-to-know-about-ai-ethics_599364
40. Alibaba Cloud Model Studio https://www.alibabacloud.com/en/product/modelstudio?_p_lc=1
41. Are Language Models Consequentialist or Deontological ... <https://arxiv.org/html/2505.21479v2>

42. Deontology and safe artificial intelligence - Springer Link <https://link.springer.com/article/10.1007/s11098-024-02174-y>
43. Concepts of Ethics and Their Application to AI - PMC <https://pmc.ncbi.nlm.nih.gov/articles/PMC7968613/>
44. Model Editing as a Double-Edged Sword: Steering Agent ... <https://arxiv.org/html/2506.20606v2>
45. Virtue, choice, and storytelling: how ethics, decision modalities ... <https://link.springer.com/article/10.1007/s10111-025-00815-8>
46. On Consequentialism and Fairness - PMC <https://pmc.ncbi.nlm.nih.gov/articles/PMC7861221/>
47. Are Language Models Consequentialist or Deontological ... <https://arxiv.org/pdf/2505.21479>
48. Advancements in Open-Source Human-Centric Neural ... <https://dl.acm.org/doi/10.1145/3703454>
49. arXiv:2312.01818v3 [cs.AI] 16 Jan 2025 <https://arxiv.org/pdf/2312.01818>
50. XITAO Contrast Color Patchwork Asymmetrical Dresses O-neck ... <https://www.aliexpress.com/item/1005009582414082.html>